

## ביולוגיה חישובית

## תרגיל 3

מגישות: הילה לוי 318905940, ונעם סרי לוי 209352624

אלגוריתם גנטי – זיהוי תבניות באמצעות רשת נוירונים

בנינו רשתות נוירונים בעזרת אלגוריתם גנטי, שילמדו את התבניות ויוכלו לנבא האם מחרוזת מסוימת מתאימה לתבנית.

**הוראות הרצה:**

עבור התוכניות buildnet0 ו-buildnet1 (קבצי exe), יש לתת כקלט את הקבצים train\_data0, test\_data0 ו-test\_data1, train\_data1 בהתאמה.

למשל, כדי להריץ את buildnet0, יש להריץ את הפקודה:

buildnet0.exe train\_data0.txt test\_data0.txt

לאחר ההרצה, יתקבלו הקבצים wnet0/1.npy בהתאמה.

עבור התוכניות runnet0 ו-runnet1 יש לתת כקלט את הקבצים wnet0, testnet0 ו-wnet1, testnet1 בהתאמה.

למשל, כדי להריץ את runnet0, יש להריץ את הפקודה:

runnet0.exe testnet0.txt wnet0.npy

לאחר ההרצה יתקבלו הקובצים classification0/1.txt בהתאמה.

תיאור הפתרון:**חלוקת הדאטה:**

חילקנו כל אחד מהקבצים wnet0/1 לקבוצת למידה ולקבוצת מבחן.

גודל קבוצת הלמידה- 15000 דוגמאות

גודל קבוצת המבחן- 5000 דוגמאות

**מבנה הרשת:**

רשת נוירונים שבנינו זהה עבור שני סוגי החוקים.

הרשת בעלת שכבה אחת של משקולות. גודל הקלט שלה הוא 16 (שהרי כל מחרוזת בינארית נתונה הינה באורך 16), וגודל הפלט שלה הוא 1 (0 או 1). בנוסף, הגדרנו מספר יחיד עבור ה-bias.

הגרלנו משקולות התחלתיים ו-bias (ערכים בין -1 ל-1 באופן אחיד).

בשלב ה-forward: הכפלת מערך המשקולות (של רשת הנוירונים) בכל דוגמא (מחרוזת), הוספת ה-bias והפעלת פונקציית האקטיבציה tanh על התוצאה שהתקבלה.

סיווג הדוגמא- אם התקבל ערך קטן מ-0 ה-output יהיה 0 (הדוגמא לא מקיימת את החוקיות), אחרת ה-output יהיה 1 (הדוגמא מקיימת את החוקיות).

האלגוריתם הגנטי למד את התבניות, ובהתאם שינה את המשקולות.

### מאפייני האלגוריתם הגנטי:

**גודל האוכלוסיה- 100.** האוכלוסיה מיוצגת בכל דור על ידי 100 רשתות נוירונים.

**מספר הדורות-** עבור  $m=0$  ו-  $m=1$  בחרנו להריץ את האלגוריתם במשך 25 דורות. זהו מספר הדורות שנדרש לאלגוריתם לרוץ על מנת ללמוד את התבנית בצורה טובה ולבנות רשת בעלת ביצועים טובים על קבוצת המבחן.

**ייצוג הפתרונות-** כל אחד מהפתרונות הוא רשת נוירונים (כפי שציינו לעיל במבנה הרשת, שכבת קלט- בגודל 16, שכבת פלט- בגודל 1) המורכבת מ:

- weights- מערך באורך 16 של משקולות
- Bias- מספר יחיד

(המשקולות וה- bias הם מספרים בין -1 ל- 1)

כל אחת מהתוכניות buildnet0 ו- buildnet1 מייצרות קובץ אחד- wnet0 ו- wnet1 בהתאמה המכילים את ערכי המשקולות וערך ה- bias שהתקבלו מהאלגוריתם אחרי כל העדכונים.

**בחירת הדור הבא-** המעבר לדור הבא התבצע באופן הבא:

1. 15% הפתרונות הטובים ביותר מהדור הקודם
2. 5% פתרונות חדשים
3. 80% פתרונות המתקבלים מביצוע פעולת ה- cross over על הפתרונות של הדור הקודם
4. מוטציות- בסיכוי 0.2 עבור כל פתרון שהתקבל מסעיף 3

**פונקציית ההערכה (fitness)-** מימשנו את הפונקציה באופן בו ככל שערך ה- fitness גבוה יותר, הפתרון טוב יותר.

עבור כל פתרון (רשת נוירונים), הפונקציה מחשבת את אחוז המחרוזות בקבוצת הלמידה שרשת הנוירונים סיווגה נכונה. כך שככל שרשת הנוירונים מסווגת מספר גדול יותר של מחרוזות בצורה נכונה, ציון ה- fitness שלה גדול יותר.

### פעולת ה- cross-over בין פתרונות:

הפתרונות שמהם מבצעים את פעולת ה- cross over: כל פתרון מהדור הקודם קיבל משקל (rank) בהתאם לערך ה- fitness שלו. הפונקציה בחרה את הפתרונות (ההורים לפעולת ה- cross over) על פי המשקלים הנ"ל, כך שככל שערך ה- fitness הוא טוב יותר, לפתרון היה יותר סיכוי להיבחר.

ה- crossover מחושב באופן הבא:

הפונקציה מגרילה באופן רנדומי ערך  $x$  בין 0 ל- 15. עבור האחוז הנבחר  $x$ , הפונקציה מחזירה שני ילדים:

ילד ראשון- מקבל את  $x$  ערכי המשקולות הראשונים מהורה 1 ואת השאר מהורה השני.  
ילד שני- מקבל את ערכי  $x$  המשקולות הראשונים מהורה 2 ואת השאר מהורה 1.

הפונקציה מגרילה מספר 0/1, אם התקבל 1:

ילד ראשון- מקבל ערך bias הזהה לשל הורה 1

ילד שני- מקבל ערך bias הזהה לשל הורה 2

אחרת:

ילד ראשון- מקבל ערך bias הזהה לשל הורה 2

ילד שני- מקבל ערך bias הזהה לשל הורה 1

כיצד מומשו מוטציות:

עבור כל הפתרון שהתקבל מפעולת ה-cross over, האלגוריתם מבצע מוטציה בסיכוי של 0.2 עבור המשקולות ועבור ה-bias (מתבצעת הגרלה עבור מוטציה למשקולות והגרלה עבור מוטציה ל-bias בנפרד).

עבור המשקולות- הפונקציה מגרילה מספר  $x$  בטווח שבין 0 ל-15. עבור המשקולות שנמצאת במקום  $x$  במערך, הפונקציה מגרילה ערך (בין 1- ל-1), מוסיפה אותו למשקולת ובכך מקטינה/ מגדילה את ערך המשקולת.

עבור ה-bias, הפונקציה מגרילה ערך (בין 1- ל-1), מוסיפה אותו לערך ה-bias ובכך מקטינה/ מגדילה את הערך הקודם.

\* הפונקציה מוודאת שהערכים נשארים בין 1- ל-1 ואם לא- מטפלת בהתאם.

האלגוריתם הגנטי רץ במשך מספר דורות (יפורט בהמשך), ובכל דור:

- מבצע ערבוב (shuffle) של קבוצת הלמידה
- עובר על 1000 פתרונות שונים (רשתות נוירונים) בכל פעם (mini batch) ומבצע:
  - o חישוב של ה-fitness של כל אחת מ-1000 הרשתות
  - o עדכון של האוכלוסיה

החוקיות מאחורי שתי התבניות:

אנחנו משערות כי החוקיות של 0חח הינה מחרוזות שבהן הסיפרה "1" מופיעה מינימום שמונה פעמים ומקסימום 12 פעמים.

אנחנו משערות כי החוקיות של 1חח הינה מחרוזות שבהן הסיפרה "1" מופיעה עד שמונה פעמים.

**ביצועי התוכנית:**

buildnet1	buildnet0	
0.998	0.994	ביצועי התוכנית על קבוצת הלמידה (ערך ה-fitness הטוב ביותר בסיום ההרצה)
0.997	0.9906	ביצועי התוכנית על קבוצת המבחן

גם עבור buildnet0 וגם עבור buildnet1, ערך ה-fitness על קבוצת הלמידה השתפר לאורך הדורות.